**Capstone Research Project Proposal**

This project studies how training language models on a variety of languages affects racial and gender bias in in these models. Specifically, I ask how the choice of languages used in training different versions of a model impacts racial and gender bias in natural language processing tasks? To answer this I examine the shifts in racial and gender bias across BERT models trained for a variety of languages.

While language models, such as BERT, have revolutionized NLP tasks, there are concerns that these models are trained to exhibit bias unintentionally. In a study done by Caliskan et al. (2017), the authors evaluated the cosine similarity between a subset of words pertaining to a variety of categories. Their tests discovered the model’s racial and gender biases. Through this study, the authors showed not only that the GloVe model exhibits these commonly held biases, but also that these biases could lead to disproportionate representation in automated decision-making processes.

I propose a similar design to Caliskan et al. in which I evaluate the cosine similarity of a subset of target words and concepts to determine if the selected models exhibit bias. All the models evaluated will be trained in both English and in at least one other language. This ensures that when running the tests for cosine similarity, no translation will be needed. The models being evaluated are available from Hugging Face. The data I will evaluate these models on will change depending on the kind of bias I want to look at.

To measure bias in these models, I test how similar a set of words is to another. For example, in the test for gender bias, I evaluate the cosine similarity between two sets of the most popular names for both men and women respectively and two sets of job labels, one consisting of STEM-related jobs and one of non-STEM-related fields. In doing so, I compare how closely related these sets of words are to each other, and whether one set of names is more closely linked to one of the two sets of job labels.

To test for biases, I will create a set of both target words and a set of concepts in my evaluations of cosine similarity. The set of target words used in each test will be consistent throughout to keep results consistent. These two sets will consist of words which are considered Pleasant (e.g. happy, kind, enjoyable) and Unpleasant (e.g. sad, angry, disgusting). I will first test for race using subsets of names that are commonly classified as European-American, African-American, Asian-American, and Latin-American, and evaluating the cosine similarity of these subsets with both the group of Pleasant and Unpleasant words. I also evaluate individual generic terms, such as Latina/Latino, African, etc. with the same group of target words. I will then test for gender bias following a similar format, in which I will gather two sets of the most common names in the US for both men and women respectively, and evaluate the cosine similarity between each set and the set of target words. With regards to gender bias, I will also evaluate the cosine similarity between these sets of names and two sets of job labels. These two sets of labels will consist of labels pertaining to STEM careers, such as engineer or data scientist, and non-STEM fields, such as journalist or teacher.

The models I will be testing are outlined in the List\_of\_Models.csv file. This csv file contains a list of 12 models, 4 of which are trained only in English and 8 multilingual models. The four English models are as follows: bert-base-uncased, roberta-base, distilbert-base-cased, and distilbert-base-uncased. These four models will serve as the control group in this experiment, as these are some of the most used variations of the BERT model, and the majority of the models I will be testing are derived from these base models. The other 8 models are: bert-base-uncased-multilingual, xlm-roberta-base, ernie-2.0-base-en, distilbert-base-multilingual-cased, stsb-xlm-r-multilingual, xlm-e, mDeBERTa-v3-base-mnli-xnli, and multilingual-MiniLMv2-L6-mnli-xnli.

To organize these models, I’ve constructed a dataset of a variety of one-hot encoded variables outlining the languages they’ve been trained in, training data, and the uses of these models, as well as if the models have been finetuned. The variables outlining the language the model was trained in are: English, Chinese, Spanish, German, French, and Multi. The Multi variable in this case represents that a model was trained in more than one language, whether that be 2 or more of the five previously listed languages, or that the model was trained in more languages than just the five listed ones. I’ve also included another variable, Number\_of\_Languages, which as the name suggests, contains the total number of languages the model was trained in through training and finetuning.

The next group of variables represents the training data that was used to train the model, or the underlying model each variation was built from. The variables are as follows: Wikipedia, BookCorpus, CommonCrawl, CC100, Ted2020, and Other. The Wikipedia variable represents the Wikipedia datasets, a collection of cleaned Wikipedia articles written in every available language. The BookCorpus variable represents the BookCorpus dataset, a collection of text from over 11,000 unpublished books scraped from the internet. The CommonCrawl dataset represents the CommonCrawl dataset, a collection of 3.15 billion webpages scraped from the internet, 46% of which are in English, with the rest being in a variety of other languages. The CC100 variable represents the CC100 corpus which was used to train XLM-R. This corpus contains data for 116 languages from the 2018 CommonCrawl snapshot. The Ted2020 variable represents the Ted2020 corpus, a collection of transcripts from nearly 4000 TED talks in 2020. Finally, the OtherData variable represents if were any other datasets used to train the model.

The intended uses of each model are also outlined in the List\_of\_Models dataset. The MLM variable represents if one of the primary uses of the model is for Masked Language Modeling, where the model is able to predict missing words in a sentence by looking at the context of the rest of the words in the sentence. The NSP variable represents the model’s ability to perform Next Sentence Prediction, where the model is given two sentences and is tasked with determining if the two sentences were following one another in whatever text they were taken from. The Sequence\_Classification variable represents if the model is capable of sequence classification, where the model is fed a sequence of data and is tasked with predicting a category for the sequence of data. The Token\_Classification variable represents if the model is capable of token classification, where the model is given a string of text and is tasked with assigning each token, in this case each word, to a category. The QnA variable represents if the model can provide a response to a question posed by a human. Lastly, the NLI variable represents if the model is capable of Natural Language Inference, in that when presented with a hypothesis, the model can determine if the hypothesis is true, false, or undetermined. If there are any other primary uses for the model which do not require finetuning, the Other\_Uses variable is used.

The final set of variables focuses on if a model was finetuned for some task(s), and the datasets that were used in the process of doing so. In the case of many of these models, they were finetuned to complete certain tasks or to be able to understand and communicate in more than one language. The finetuning variables used in this dataset represent the names of different finetuning datasets that a model uses. The variables in this section are as follows: XNLI, XQuAD, GLUE, SuperGLUE, MNLI, and OtherFT. All of these variables, apart from OtherFT, represent a specific well-known and trusted dataset used in finetuning language models. The OtherFT variable represents if the model was finetuned on any other dataset not listed previously.

Sources

Caliskan, Aylin, et al. “Semantics Derived Automatically from Language Corpora Contain Human-like Biases.” *Papers With Code*, 25 Aug. 2016, paperswithcode.com/paper/semantics-derived-automatically-from-language.

“Bert-Base-Uncased · Hugging Face.” *Bert-Base-Uncased · Hugging Face*, huggingface.co/bert-base-uncased. Accessed 17 May 2023.